CITD@UIT at SemEval-2026 Task 4: Structured Reasoning and Metric Specialization for Narrative Similarity

Thach Ngoc Nguyen[1,2], Duc-Vu Nguyen[1,2], Dang Van Thin[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
25210035@ms.uit.edu.vn {vund, thindv}@uit.edu.vn

Abstract

We present a synergistic dual-track approach for SemEval-2026 Task 4 on narrative similarity, covering Track A (triple-wise classification) and Track B (narrative representation) through failure-driven data enrichment. The shared task received 71 final submissions from 46 teams across its two tracks. For Track A, we explore three reasoning strategies: hybrid Cross-Encoder–LLM arbitration (66.5% dev), DSPy-based component-wise decomposition (68.0% dev), and a multi-stage pairwise reasoning pipeline with enforced moral agency hierarchies, where the final Gemini 2.5 Pro/Flash system achieves **77.39%** on development and **69.25%** on test data, ranking 17th among 46 participating teams in the official evaluation. For Track B, we propose **BGE-M3 (LoRA)**, an instruction-guided dense representation model trained with Multiple Negatives Ranking Loss (MNRL); since Track B provides only unlabeled story instances, we specialize the embedding space using adversarial samples synthesized from Track A failure cases, achieving **68.75%** in the official evaluation and ranking 6th among 26 participating teams. Our analysis shows that narrative similarity depends more on outcome alignment and moral trajectory than lexical overlap, highlighting the complementary roles of explicit reasoning and task-specific metric-space specialization.

1 Introduction

SemEval-2026 Task 4 (Hatzel et al., 2026) challenges systems to model narrative similarity beyond surface-level lexical overlap. The task comprises two complementary tracks: Track A requires selecting which of two candidate stories is narratively closer to an anchor story, while Track B requires producing dense vector representations whose geometry aligns with narrative similarity judgments.

Surface-level embeddings frequently overestimate similarity when stories share entities or settings but diverge in moral trajectory. For example, two stories may involve astronauts, yet differ fundamentally in their ethical framing. This "lexical trap" phenomenon motivates our dual approach: structured reasoning for discriminative classification (Track A) and metric-space restructuring for narrative representation (Track B).

Unlike traditional Semantic Textual Similarity (STS), narrative similarity requires modeling temporally extended causal structures and protagonist-level intention alignment. This introduces a *structural invariance problem*: similarity must be sensitive to causal transformations but invariant to surface lexical features. We hypothesize that narrative similarity lies in a structured manifold governed primarily by **moral agency** and **outcome alignment**. This motivates a methodology that combines symbolic-like decomposition (Track A) and metric geometry restructuring (Track B).

To address these challenges, we employ two distinct yet synergistic methodologies designed to distill raw narratives into their essential structural components. For Track A, we progressively transition from surface-level lexical modeling to structured narrative reasoning, utilizing DSPy to enforce rigorous Chain-of-Thought (CoT) decomposition. For Track B, we reformulate narrative representation as an instruction-guided dense retrieval task. We specialize the embedding space using a **BGE-M3 (LoRA)** model, fine-tuned via cross-task supervision derived from the reasoning signals and failure cases encountered in Track A.

The main contributions of this work are as follows:

1. We demonstrate that structured pairwise reasoning with an enforced moral agency hierarchy substantially outperforms both lexical-based arbitration and independent scoring paradigms, achieving 77.39% on dev and 69.25% on test for Track A.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1153–1162

2. We introduce a cross-task metric specialization approach using **BGE-M3 (LoRA)** that transfers supervision from Track A to Track B through failure-driven adversarial augmentation, achieving 68.75% in official track B.

3. We provide empirical evidence that narrative similarity is primarily governed by outcome polarity and moral agency rather than lexical overlap, and that classification and retrieval objectives induce geometrically divergent representations.

2 Related Work

Semantic vs. Narrative Similarity. Sentence-level similarity models such as Sentence-BERT (Reimers and Gurevych, 2019) have achieved strong results on STS benchmarks but struggle with narrative structures that require reasoning about plot arcs, character archetypes, and causal logic beyond topical overlap.

Programmatic LLM Reasoning. DSPy (Khattab et al., 2024) treats LLM prompts as tunable modules within a pipeline, enabling multi-stage strategies such as Chain-of-Thought and self-reflection. This programmatic paradigm is essential for narrative analysis, where direct prompting is easily misled by surface-level lexical cues. Advanced reasoning models like Gemini 2.5 (Comanici et al., 2025) further enable extraction of abstract structural dimensions.

Long-Context Dense Retrieval. BGE-M3 (Chen et al., 2024) supports sequences up to 8,192 tokens, a critical capability for narrative representation where plot trajectories span thousands of tokens. Combined with instruction-guided encoding, dense retrieval models can be directed toward structural semantics rather than keyword matching.

Parameter-Efficient Fine-Tuning. LoRA (Hu et al., 2022) enables domain specialization by introducing low-rank adapters while freezing backbone weights, avoiding catastrophic forgetting. This approach is well-suited for adapting general-purpose embeddings to narrative-specific similarity tasks with limited supervision.

Narrative Structure in NLP. Classical narrative analysis frameworks such as Plot Units (Lehnert, 1981) and Narrative Event Chains (Chambers and Jurafsky, 2008) formalize stories as structured sequences of causally-linked events. More recently, Story Intention Graphs (Finlayson, 2012) provide symbolic representations of character goals and outcomes. Our work draws on these traditions by decomposing narratives into structural slots (theme, action, outcome) while leveraging LLMs as flexible parsers rather than hand-crafted rules.

3 Methodology

Our methodology is divided into two parts: a series of evolving reasoning paradigms for the triple-wise classification in Track A, and a narrative-aware structural encoding framework for the embedding task in Track B.

3.1 Formalization of Narrative Similarity

We formalize narrative similarity between an anchor X and a candidate A as a weighted combination of the three core components defined by the task:

$$S(X, A) = \alpha S_{\text{theme}} + \beta S_{\text{action}} + \gamma S_{\text{outcome}} \quad (1)$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$. In our implementation, we prioritize the **Outcome** (γ) as the non-negotiable discriminator, enforcing a hierarchy where $\gamma > \alpha, \beta$.

In addition to the standard narrative components, an empirical analysis of misclassified instances suggests that superficial lexical traps and sensitive terminology frequently obfuscate the underlying narrative arc, often triggering restrictive safety filters in Gemini 2.5 Pro that result in null outputs . To mitigate these systemic constraints, we formalize Moral Agency as an auxiliary evaluative dimension. This framework enables the system to extract the latent intentionality and ethical deliberations of the protagonist, providing a universally applicable lens that maintains consistency across the entire dataset without compromising the core narrative essence. By prioritizing these intrinsic features, our approach neutralizes the influence of spurious lexical traps and enhances plot clarity. Furthermore, this abstraction ensures high thematic fidelity while successfully circumventing API-level systemic constraints, allowing for robust processing of diverse and sensitive narrative content.

3.2 System Overview

Figure 1 illustrates the technical architecture of our dual-track approach. Track A focuses on struc-

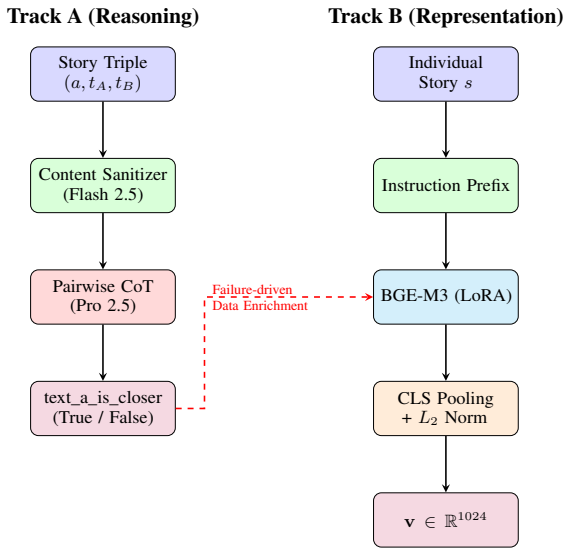

Figure 1: System overview of our synergistic approach.

tured reasoning, while Track B focuses on metric specialization.

Track A employs a sequential reasoning pipeline where **Gemini 2.5 Flash** sanitizes input narratives to ensure noise reduction and policy compliance before deep reasoning, **Gemini 2.5 Pro** predicts `text_a_is_closer`.

Track B utilizes an instruction-guided **BGE-M3 (LoRA)** model, trained on **3,223 samples** synthesized by combining the official Track A development set, the organizer-provided synthetic classification data, and **1,126 newly generated adversarial examples** modeled after Track A's specific failure modes, specifically leveraging Lexical Anchors and Narrative Pivots to decouple surface motifs from core logic.

3.3 Track A: Structured Reasoning

We explored three distinct approaches for Track A, where each subsequent iteration emerged from unsuccessful attempts to improve the performance of its predecessor, leading to alternative insights into narrative modeling.

3.3.1 Approach 1: Hybrid Neural–Symbolic Ensemble

In our initial attempt, we developed a multi-layered ensemble to leverage the complementary strengths of encoder-based lexical precision and LLM-based reasoning. This system integrated three components: (1) *DeBERTa-v3-base* as a Cross-Encoder for NLI-based alignment, (2) *Gemini 2.5 Flash* as the primary reasoning engine, and (3) *Gemini 2.5 Pro* as the final symbolic arbiter.

The workflow followed a voting consensus. *DeBERTa* provided lexical signals, while *Gemini 2.5 Flash* executed two symmetric runs with A/B position swapping to mitigate positional bias. An **Ethical Arbitration** module, powered by *Gemini 2.5 Pro*, was triggered only when a conflict was detected, either between the Cross-Encoder and the LLM or when the Flash model's swapped runs were inconsistent.

This module is designated "Ethical Arbitration" because it is specifically tasked with resolving conflicts by prioritizing the **Course of Action** over deceptive surface-level motifs. The arbiter resolves these conflicts using a strict hierarchy where the Course of Action is explicitly clarified through: (1) **Moral Agency & Conflict** (providing the ethical intent behind actions), (2) **Plot Arc** (mapping the structural progression), and (3) **Outcome Tone**. By focusing on the "ethical" deliberations of the protagonist, the system ensures that the narrative's causal backbone remains clear even when explicit keywords are sanitized or trigger safety filters.

Despite this robust tri-model architecture, it achieved an accuracy of only **66.5%** on the development set, as the complexity of reconciling divergent signals from three different models introduced unforeseen noise.

3.3.2 Approach 2: Component-Based Scoring via Structural Decomposition

To improve results, we moved toward a more structured representation. Using the *DSPy* framework, we decomposed each story into five core narrative dimensions: *Genre & Setting, Mood, Archetypes, Plot Structure*, and *Outcome*.

In this paradigm, candidates were evaluated via Absolute Scoring, where each story was assigned a discrete similarity score $s \in \{0, 10, \ldots, 100\}$. A *Self-Reflection* module audited these scores for consistency. While this increased accuracy to **68.0%** on the development set (**66.75%** on the test set), the system struggled with fine-grained relative nuances. This performance gap highlights the inherent difficulty of absolute rating, which is highly susceptible to calibration noise, since assigning consistent numerical scores across diverse samples is significantly more challenging for LLMs than identifying the relative "narrative delta" between two candidates in a direct comparison.

3.3.3 Approach 3: Symmetric Pairwise Reasoning (Final Optimized System)

Refining our insights from previous attempts, we abandoned the complexity of hybrid ensembles and absolute scoring in favor of a Relative Comparison philosophy. This approach focused on the superior discriminative power of advanced LLMs in a direct competitive setting, achieving our peak performance of **77.39%** on the development set (**69.25%** on the test set).

- **Narrative Sanitization (Gemini 2.5 Flash):** We utilized the Flash model to distill raw summaries into structural representations, stripping away non-narrative noise. Crucially, this module re-writes the content to be AI-safety compliant while **strictly preserving the narrative DNA and moral agency**. By focusing on the **Course of Action** rather than specific sensitive terminology, this process prevents "null responses" without distorting the story's fundamental ethical trajectory. While this may occasionally result in "thematic over-abstraction", it ensures a robust and consistent input for the subsequent reasoning stage.

- **Symmetric Pairwise CoT (Gemini 2.5 Pro):** The core reasoning was performed by *Gemini 2.5 Pro*. To eliminate **Positional Bias**, we executed pairwise comparisons for both orders $(Anchor, A, B)$ and $(Anchor, B, A)$. We enforced a **Chain-of-Thought (CoT)** prompt, requiring the model to generate a step-by-step analysis of the Course of Action, specifically examining the moral agency (the intentionality behind character choices) and causal progression. Only consistent results across both permutations were accepted, ensuring the decision was rooted in stable narrative logic rather than input sequence.

3.4 Track B: Metric Specialization for Narrative Similarity

For Track B, we focused on specializing the embedding space to prioritize deep narrative structures over surface-level lexical overlaps. We developed two distinct systems representing different philosophies: architectural regularization via auxiliary narrative states and failure-informed contrastive learning.

3.4.1 Approach 1: Structural Regularization via Outcome Classification

Our first system, **NADSE** (Narrative-Aware Dense Structural Encoding), aimed to inject explicit structural knowledge into the latent space. This approach utilized a *BAAI/bge-large-en-v1.5* backbone enhanced with **Low-Rank Adaptation (LoRA)** ($r = 16, \alpha = 32$).

To make the embeddings "narrative-aware," we integrated an auxiliary **Outcome State Classifier** based on *cross-encoder/nli-deberta-v3-xsmall*. This component was trained to categorize stories into three fundamental narrative states: SUCCESS, FAILURE, or AMBIGUOUS. These signals served as a mechanism for **Latent Space Regularization**, aligning with our formal hierarchy (Section 3.1) where Outcome (γ) is the primary discriminator. By explicitly identifying the resolution of a story, the model sought to force apart narratives with diametrically opposed outcomes in the metric space, even when they shared significant lexical overlap. This system achieved an accuracy of **61.25% on the test set**.

3.4.2 Approach 2: Failure-Informed Contrastive Learning

Our final and best-performing system for Track B shifts toward a data-driven strategy informed by systematic model failures. To implement this, we transitioned to the *BAAI/bge-m3* backbone, leveraging its superior multi-stage retrieval capabilities and enhanced sensitivity to instruction-guided encoding compared to the BGE-Large model used in Approach 1. We designate this specialized configuration as **N-BGE-LoRA** (Narrative-BGE-LoRA). This approach utilizes **Low-Rank Adaptation (LoRA)** ($r = 16, \alpha = 32$) to specialize the representation space for plot-centric similarity and is optimized using the **Multiple Negatives Ranking Loss (MNRL)**:

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(a_i, p_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(a_i, p_j)/\tau)} \quad (2)$$

The core innovation lies in its **Failure-Informed Data Synthesis** process, creating a training set of **3,223 samples**:

- **Gold-Standard Triplets (200 samples):** Official development data (*dev_track_a.jsonl*).

- **Cleaned Synthetic Data (1,897 samples):** Sourced from the official synthetic corpus,

with three null entries removed to ensure signal integrity.

- **Adversarial "Trap" Samples (1,126 samples):** These samples were synthesized through the Gemini 2.5 web-based conversational interface. We opted for the interactive web platform over API-based generation to support iterative prompt refinement and manual quality checkpoints during narrative synthesis. Using a structured debugging-oriented prompt strategy, detailed in Appendix B, we used misclassified triples from Track A as seeds to synthesize two types of adversarial variants:
  - *LANP (Lexical Anchors, Narrative Pivots):* Termed "**The Trap**", these samples aggressively reuse the setting and keywords (Lexical Anchors) while inverting the **Course of Action** and **Outcome** (Narrative Pivots). This forces the model to move beyond the "Setting Trap" and "Keyword Trap."
  - *NALP (Narrative Anchors, Lexical Pivots):* Termed "**The Disguise**", these samples preserve the core narrative arc and outcome (Narrative Anchors) while transplanting them into radically different lexical domains (Lexical Pivots) to encourage domain-invariant structural learning.

Instruction-Guided Encoding: We prefix narratives with: *"Represent this story for narrative similarity retrieval focusing on plot structure and outcome:"*. This acts as a **Global Attention Filter**, biasing the transformer's self-attention to prioritize causal transitions while penalizing idiosyncratic entities. Parameter-efficient specialization is achieved via LoRA adapters: $W' = W + BA$, where $r = 16, \alpha = 32$. This system achieved our highest Track B performance with an accuracy of **68.75% on the test set**.

4 Experiments

4.1 Setup

Computational Resources. We optimized our hardware allocation based on task requirements. Track A experiments, including Cross-Encoder inference and LLM-based reasoning, were conducted on **NVIDIA T4** GPUs (16 GB) via Google Colab and Kaggle. For the training-intensive Track B, we utilized an **NVIDIA A100** GPU (40 GB) on Google Colab. Fine-tuning our N-BGE-LoRA model on 3,223 triplets with an 8,192-token context window was completed in approximately **36 minutes** (2 epochs).

Track A. The pipeline utilizes *DeBERTa-v3-base* (He et al., 2021) as an NLI Cross-Encoder to provide lexical signals. For narrative reasoning, we employ **Gemini 2.5 Flash** (for sanitization and initial CoT) and **Gemini 2.5 Pro** (for complex pairwise reasoning and conflict arbitration). For Approach 1 (Hybrid Arbitration), we use structured JSON mode with $\tau = 0.0$ for deterministic outputs. In Approach 3 (DSPy system), Gemini 2.5 Flash uses $\tau = 0.2$ for sanitization, and Gemini 2.5 Pro uses $\tau = 0.5$ for pairwise comparison via DSPy (Khattab et al., 2024). The system's performance is evaluated using Accuracy.

Track B. Full hyperparameters are listed in Table 1.

Parameter	Value
Backbone	BGE-M3
Max sequence length	8,192 tokens
LoRA rank (r)	16
LoRA scaling (α)	32
LoRA dropout	0.1
LoRA target modules	Q, K, V, Dense, Proj
Pooling	CLS + L_2 Norm
Loss	MNRL (Henderson et al., 2017)
Temperature (τ)	0.05
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Learning rate	1×10^{-4}
Batch size	2
Epochs	2
Warmup steps	100
Precision	FP16 (mixed)
Gradient checkpointing	Enabled
Training corpus	3,223 triplets
Embedding dimension	1,024

Table 1: Track B hyperparameters for N-BGE-LoRA.

4.2 Track A Results

The performance of our three iterative approaches for Track A is summarized in Table 2. The results demonstrate a clear upward trend as we transitioned from hybrid ensemble heuristics to structured pairwise reasoning.

Analysis of Methodology Evolution. The progression across the three methods validates our strategic shift in framing narrative similarity:

Approach	Paradigm	Dev Acc.	Test Acc.
Approach 1	Hybrid Ensemble	66.50%	-
Approach 2	Component Scoring	68.00%	66.75%
Approach 3	**Symmetric Pairwise**	**77.39%**	**69.25%**

Table 2: Evolution of Track A systems. Approach 3 demonstrates that relative narrative ranking significantly outperforms absolute component scoring by mitigating calibration noise.

- **From Conflict to Consensus (Approach 1 vs. Approach 3):** Approach 1 attempted to reconcile signals from a Cross-Encoder and two LLM runs. However, the complexity of the arbitration logic often introduced noise when the encoder's lexical bias conflicted with the LLM's structural analysis. By centering the pipeline on symmetric LLM reasoning in Approach 3, we achieved a significant gain of **+10.89%** on the development set.

- **Relative vs. Absolute Judgment (Approach 2 vs. Approach 3):** The transition from Approach 2 to Approach 3 represents the most substantial internal improvement (**+9.39%** on dev). We observe that Approach 2's reliance on a discrete 10-point scale (e.g., $\{0, 10, \ldots, 100\}$) is inherently prone to *calibration noise* and subjective bias. In such a regime, the distinction between adjacent tiers (e.g., a score of 60 vs. 70) is often fuzzy and highly sensitive to minor linguistic variations in the story summaries. A marginal miscalibration of just 10 points, which is common in absolute LLM rating, is sufficient to erroneously flip the final comparative decision. Approach 3 mitigates this issue by shifting from a *rating* task to a *ranking* task, leveraging the LLM's stronger ability to perform direct, fine-grained relative discrimination, namely the *narrative delta*, within a single shared context window. This removes the need for arbitrary numerical mapping.

Robustness via Sanitization. A critical factor in the stability of Approach 3 (achieving **69.25%** on the test set) was the content sanitization step. By rewriting raw Wikipedia plots into safety-compliant summaries, we mitigated API refusal triggers and ensured that the reasoning engine focused exclusively on the structural "DNA" of the narrative rather than sensitive surface-level terminology.

4.3 Track B Results

The evaluation of our narrative representation models for Track B is presented in Table 3. We compare our initial structural regularization framework (NADSE) against our final specialized model, N-BGE-LoRA.

Approach	Backbone	Key Strategy	Test Score
Approach 1	BGE-large	Outcome Reg.	61.25%
Approach 2	**BGE-M3**	**Failure-Informed**	**68.75%**

Table 3: Performance of Track B embedding models. The integration of adversarial "Trap" samples significantly improved structural robustness.

Analysis of Model Evolution. The transition from Approach 1 to Approach 2 highlights the synergistic effect of expanded context capacity and specialized, failure-informed supervision:

- **Overcoming Lexical Gravity:** Approach 1 (NADSE) utilized a *BGE-large* backbone with an explicit gating mechanism for *Setup, Actions,* and *Outcome*. While theoretically sound, this model was highly susceptible to lexical gravity, a phenomenon where surface-level keyword overlaps (e.g., shared settings or character names) exert a disproportionate "pull" on the embedding space and overshadow underlying structural differences. Furthermore, the 512-token limit led to the truncation of critical narrative pivots (γ). By adopting BGE-M3 with an 8,192-token window, N-BGE-LoRA captured the full causal arc more effectively.

- **Synergy of Data Enrichment and Model Capacity:** The performance gain (**+7.50%**) stems from the combined effect of a more capable backbone and a curated training corpus. In Approach 2, we fine-tuned the model using **Multiple Negatives Ranking Loss (MNRL)** on a strategically aggregated dataset of **3,223 triplets**, comprising:

  1. **200 gold-standard triples** from the official Track A development set.
  2. **1,897 synthetic triples** filtered from the original 1,900-sample synthetic corpus provided by the organizers (after removing 3 null entries to ensure signal quality).

3. **1,126 adversarial "trap" samples** synthesized via the Gemini 2.5 web interface. These were specifically designed to target the lexical and setting biases identified in our Track A error analysis.

This curated mixture forced the embedding space to decouple plot structure from surface vocabulary, utilizing the backbone's increased capacity to prioritize deep causal transitions and moral agency.

Rationale for Data Selection. Although Track B focuses on individual story representations, we utilized the labeled triplets from Track A as a *cross-task supervisory signal* for training. This decision was driven by the requirements of contrastive metric learning: while the dedicated Track B data consists of unlabeled instances, the Track A dataset provides explicit relative similarity judgments (Anchor, Positive, Negative). These labels are essential for optimizing the **Multiple Negatives Ranking Loss (MNRL)**, enabling the model to learn a manifold where narrative arcs are geometrically aligned according to human-annotated logic.

Final Submission Details. Our official entry (N-BGE-LoRA) employs **LoRA** adapters ($r = 16, \alpha = 32$) and an instruction-guided prefix: *"Represent this story for narrative similarity retrieval focusing on plot structure and outcome:"*. The final representations are 1,024-dimensional vectors produced via CLS pooling with L_2 normalization, ensuring a robust metric space for cosine similarity judgments.

Generalization to Pairwise Discrimination. To further investigate the robustness of our learned representations, we evaluated the trained N-BGE-LoRA model on the Track A test set. It is important to note that while the model was trained on a combination of gold-standard dev triples and synthetic data, it had no prior exposure to the Track A test instances, ensuring a fair assessment of its generalization capabilities. We computed the cosine similarity S_{cos} between the anchor embedding $\mathbf{e}_{anc}$ and the two candidate embeddings $\mathbf{e}_A, \mathbf{e}_B$. The choice was made by selecting the candidate with the higher similarity score:

$$\hat{y} = \mathbb{I}(S_{cos}(\mathbf{e}_{anc}, \mathbf{e}_A) > S_{cos}(\mathbf{e}_{anc}, \mathbf{e}_B))$$

Using this embedding-based inference, the model achieved an accuracy of **59.00%** on the Track A test set. When compared to our official Track A reasoning-based system (which achieved **69.25%**), we observe a performance gap of **10.25%**.

This result highlights a fundamental distinction between *narrative representation* and *narrative reasoning*. While our contrastive training successfully mapped stories into a manifold that captures broad narrative similarities (as evidenced by the **68.75%** score on Track B), the task of binary discrimination in Track A often hinges on subtle causal pivots and moral nuances. These fine-grained details are more effectively captured by the explicit, multi-stage reasoning of our Track A system than by the global aggregate vectors of the N-BGE-LoRA model.

4.4 Error Analysis

Our qualitative examination of the misclassified instances in Track A reveals three primary failure modes, suggesting that deep narrative understanding requires capturing the subtle interplay between character intent and causal transitions, rather than just high-level structural extraction. Detailed case studies for each mode are presented in the appendix tables, specifically Tables 4–6 in Appendix A.

- **Lexical Gravity:** This represents a persistent susceptibility to "Entity Traps." **Lexical Gravity** occurs when the high semantic weight of specific motifs, such as "news reporter" or "murder mystery," exerts a disproportionate "pull" on the model's attention. This causes the system to ignore fundamentally divergent causal chains (γ) simply because the stories share a dominant lexical domain. *See Table 4 in Appendix A for the "Investigative Entity" trap.*

- **Moral Agency Rigidity:** The system occasionally fails to navigate ethical "gray zones" or non-linear character development. This manifests as a bias toward **surface settings** (e.g., academic or workplace environments) rather than the **moral or intellectual transformation** of the protagonist. The models struggle when a character's growth arc does not follow a conventional positive/negative trajectory, often exacerbated by safety sanitization. *See Table 5 in Appendix A for a comparison involving sensitive boundary-crossing.*

- **Thematic Over-abstraction:** The content sanitization step, while essential for safety,

occasionally strips away unique thematic nuances. This results in a "generic trope bias" where the model prioritizes high-level dynamics (e.g., standard crime-flight motifs) over specific, redemptive relationship pivots. In these cases, the model's latent representation collapses multiple distinct narratives into a single prototypical category. *See Table 6 in Appendix A for the "Redemptive Partnership" vs. "Criminal Flight" trade-off.*

5 Discussion

The results across both tracks reveal several broader insights into the nature of narrative similarity and the different roles played by sanitization, reasoning, representation learning, and adversarial tuning.

Sanitization vs. Reasoning: Unpacking the Performance Driver. To address concerns regarding whether the performance peak is driven primarily by the sanitization step, Gemini 2.5 Flash, we compare Approach 2 (Component-based Scoring) and Approach 3 (Symmetric Pairwise). Both methodologies utilize structural decomposition and rewriting, a form of sanitization, via Gemini 2.5 Flash. However, Approach 3 demonstrates a significant performance gain of **+9.39%** over Approach 2 on the development set. This substantial delta indicates that sanitization serves primarily as a preprocessing layer to ensure API safety compliance and reduce surface-level lexical noise. The true catalyst for accuracy is the multi-stage symmetric pairwise reasoning within a shared context window, which effectively eliminates *calibration noise*, namely the difficulty of assigning consistent absolute scores, and enables the model to focus on the fine-grained "narrative delta" between candidates.

The Representation vs. Reasoning Gap. The **10.25% performance gap** observed on the Track A test set (69.25% for reasoning vs. 59.00% for embedding) suggests that global aggregate vectors still lack the "resolution" required for fine-grained narrative discrimination. While Model B (N-BGE-LoRA) effectively maps stories into a general semantic manifold, it cannot yet emulate the explicit, multi-stage causal verification that the Track A system performs. This reinforces the need for **hybrid neural-symbolic approaches** in narrative NLP.

Warping the Metric Space via Adversarial Tuning. Our experiments in Track B demonstrate that adversarial triplet tuning does more than just increase accuracy; it effectively warps the embedding geometry. By penalizing reliance on lexical shortcuts (the "Trap" samples), the model is forced to neutralize Lexical Gravity and learn a *narrative-specific metric* where outcome alignment and moral agency carry more weight than entity overlap.

Toward a Manifold Hypothesis for Stories. We hypothesize that narrative similarity exists on a low-dimensional manifold defined by intentionality and consequence. The success of shifting from absolute "Rating" (Approach 2) to relative "Ranking" (Approach 3) in Track A suggests that LLMs are better at navigating this manifold through comparative deltas rather than arbitrary numerical mapping. Future work should investigate whether explicit "Story Intention Graphs" can be used to further bridge this gap.

6 Conclusion

We presented a unified dual-track framework for narrative similarity modeling in SemEval-2026 Task 4, showing that narrative similarity depends on structured reasoning over abstract narrative components rather than simple semantic overlap. For Track A, our **symmetric multi-stage reasoning pipeline** achieved **77.39%** on development and **69.25%** on the test set. By combining content sanitization with direct pairwise ranking, the pipeline outperformed absolute scoring methods and complex ensemble heuristics. For Track B, our N-BGE-LoRA contrastive model reached **68.75%** test accuracy by specializing the metric space through adversarial triplet tuning on 1,126 synthesized "trap" samples, demonstrating the value of targeted data enrichment for representation learning. Overall, our findings yield three key insights: (1) narrative similarity is primarily governed by **moral agency and outcome alignment**; (2) structured ranking effectively mitigates LLM calibration noise; and (3) adversarial training is crucial for neutralizing Lexical Gravity, namely surface-level "Keyword Traps." Future work will focus on integrating **Explicit Narrative Graphs** to further bridge the gap between neural flexibility and symbolic interpretability.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund. We thank the anonymous

reviewers for their time and helpful suggestions that improved the quality of the paper.

References

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2502.10001.

Mark Alan Finlayson. 2012. *Learning Narrative Structure from Annotated Folktales*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA. Advisor: Patrick H. Winston.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2025 task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Wendy G. Lehnert. 1981. Plot units and narrative summarization. *Cogn. Sci.*, 5:293–331.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

A Qualitative Case Studies

This section presents three representative case studies from our Track A error analysis, summarized in Tables 4–6. Each case illustrates a distinct failure mode discussed in Section 4.4, showing how current embedding models and LLM-based reasoning can be misled by surface-level cues rather than deeper narrative structure. Table 4 demonstrates a case of lexical gravity, where dense investigative vocabulary pulls the model toward the wrong candidate despite weaker plot-level similarity. Table 5 shows how moral-agency distinctions can be blurred when sensitive or transgressive relationships are overly neutralized, causing the model to confuse fundamentally different interpersonal dynamics. Finally, Table 6 illustrates thematic over-abstraction, where the model collapses two narratives into a generic action trope while overlooking the more important structural role of redemption, human connection, and mercy.

Case 1: Lexical Gravity (The "Investigative Entity" Trap)
Anchor. A janitor, Daryll, is an avid fan of a news reporter, Toni. After a murder occurs in his building, Toni suspects that Daryll knows something. Daryll plays along to stay close to her, which leads the killers to believe that he has vital information and places both of them in danger. **Text A (Distractor).** A truck driver is murdered by the Mafia. A police captain, Bellodi, investigates irregularities and corruption, struggling against an honor system in which witnesses withhold information. **Text B (Ground Truth).** A woman, Dédée, and her pimp flee to Antwerp. She meets a sympathetic captain, and they fall in love. The pimp kills the captain, and Dédée eventually takes revenge. **Model Prediction.** Text A. **Ground Truth.** Text B. **Analysis.** The model succumbs to **lexical gravity**. Although the Anchor shares a romantic-crime arc with Text B, the system is pulled toward Text A by the high density of investigative terms, such as *murdered*, *suspected*, *witnesses*, *information*, and *false accusations*. These lexical signals outweigh the deeper structural similarity, namely a romance that escalates into fatal criminal conflict.

Table 4: Case 1 – Lexical gravity causes the model to prefer the investigative distractor over the structurally closer romantic-crime narrative.

Case 2: Moral Agency Rigidity (The "Safety Neutralization" Bias) — #110 in the test set
Anchor. A man with an attractive wife develops feelings for a warm-hearted secretary because of her kindness. This relationship changes his life. **Text A (Ground Truth).** A plot involving adult-themed infidelity and the emotional reassessment of a marriage. **Text B (Distractor).** A teacher builds a connection with a 16-year-old student. The student manipulates situations and crosses social boundaries, leading to consequences for the teacher. **Model Prediction.** Text B. **Ground Truth.** Text A. **Analysis.** This case illustrates **moral agency rigidity**, amplified by safety sanitization. The Anchor and Text A share a grounded arc of marital infidelity and emotional reassessment, but the model selects Text B because of superficial overlap in professional or academic settings and the vague notion of a "complex relationship." By neutralizing the transgressive teacher–student power dynamic in Text B, the model fails to distinguish between a failing marriage and a manipulative predator–victim structure.

Table 5: Case 2 – Moral agency rigidity leads the model to overlook the distinction between marital infidelity and a transgressive predator–victim dynamic.

Case 3: Thematic Over-abstraction (The "Generic Trope" Collapse)
Anchor. An ex-convict, Lucas, is taken hostage by an incompetent novice robber, Ned. They form an unlikely partnership to save Ned's ill daughter and flee to Canada. **Text A (Distractor).** A man, Mux, documents his tryst with justice, but eventually shoots his girlfriend and flees to Italy with a simple-minded colleague. **Text B (Ground Truth).** Three men plot a heist. A commissioner recognizes one of them as the man who saved his life in the army years earlier, leading to a moment of recognition and mercy. **Model Prediction.** Text A. **Ground Truth.** Text B. **Analysis.** The system demonstrates **thematic over-abstraction**. It collapses the Anchor and Text A into the generic trope of "flight with a simple-minded partner." In doing so, it ignores the Anchor's **redemptive agency**, where flight is motivated by the desire to save a child, and contrasts it insufficiently with Text A, where flight is used to conceal a murder. The model therefore prioritizes surface-level action patterns over the deeper structure of human connection and mercy shared by the Anchor and Text B.

Table 6: Case 3 – Thematic over-abstraction causes the model to favor a shared flight trope rather than the deeper structure of redemption and mercy.

B Adversarial Synthesis Prompt

To generate high-quality adversarial triplets, we used a structured LLM prompt with a specialized persona and explicit instructions targeting two specific failure modes: Lexical Bias and Length Bias. Although the prompt fixed `text_a_is_closer` as `true` during generation for consistency and easier validation, we later balanced the labels through randomized post-processing by swapping `text_a` and `text_b` in 50% of the generated samples and updating the labels accordingly. The raw prompt used to synthesize the 1,126 "Trap" samples is shown below for reproducibility and further qualitative inspection:

```
Role: You are a Senior Data Scientist
    specializing in NLP, specifically in
    debugging DeBERTa-based models for Narrative
     Similarity. I have analyzed the model's
    errors and identified specific weaknesses (
    Lexical Bias and Length Bias). You can refer
     to following attachment file (
    incorrect_predictions).

Task: Generate 10 high-quality training triples
    (Anchor, Text A, Text B) designed to break
    the model's reliance on keywords and force
    it to learn semantic structure.

Input Analysis (Targeted Weaknesses):
- The "Setting Trap": The model incorrectly
    thinks stories are similar just because they
     happen in the same place (e.g., Hospital,
    Prison, Wild West) or involve the same job (
    e.g., Detective, Nurse).
- The "Keyword Trap": The model over-relies on
    specific nouns (e.g., "River", "Money", "
    Explosion").
- Length Bias: The model fails on longer,
    complex narratives (200+ words).

Requirements for Each Example:
1. Anchor (Complex Narrative):
   - Create a story summary (150-250 words).
    Make it complex with multiple turns or a
    specific abstract theme (e.g., "Redemption
    through sacrifice," "The corruption of
    innocence").
   - Include specific settings (e.g., A rainy
    noir city, a sci-fi spaceship) and tangible
    objects.
2. Text A (The Semantic Match - "The Disguise"):
   - Logic: MUST match the Anchor's Theme, Plot
    Structure, and Outcome perfectly.
   - Vocabulary: MUST use a completely different
     setting and vocabulary.
3. Text B (The Hard Negative - "The Trap"):
   - Logic: MUST have a completely different
    Theme or Outcome (e.g., Tragedy vs. Comedy,
    Success vs. Failure).
   - Vocabulary: MUST aggressively reuse the
    setting, keywords, and character archetypes
    from the Anchor.

Label: text_a_is_closer is always true.

Output Format: Provide strictly a JSON list
    containing 10 objects.

JSON Example:
[
  {
    "anchor": "...",
    "text_a": "...",
    "text_b": "...",
    "text_a_is_closer": true
  }
]
```